DOCUMENT RESUME

ED 442 858 TM 031 278

AUTHOR Lee, Guemin
TITLE Estimating Reliability and Standard Error of Measurement for Complex Reading Comprehension Tests under Generalizability Theory Models.
PUB DATE 2000-04-24
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Error of Measurement; *Estimation (Mathematics); *Generalizability Theory; *Models; *Reading Comprehension; Reading Tests; *Scores

ABSTRACT

The purpose of this study was to investigate the relative appropriateness of several procedures for estimating reliability and standard errors of measurement of complex reading comprehension tests. Seven generalizability theory models were conceptualized by incorporating one or several factors of items, passages, themes, contents, and types of passages as sources of score variation. Results indicate that generalizability (reliability-like) coefficients for multivariate generalizability theory models incorporating "contents" and "types of passages" are close to coefficient alpha and, in contrast, incorporating "passages" and "themes" within univariate generalizability theory models produce non-negligible differences in reliability from coefficient alpha. This suggests that passages and themes be considered in evaluating the reliability of test scores for complex reading comprehension tests. (Contains 1 figure, 7 tables, and 15 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 442 858

# Estimating Reliability and Standard Error of Measurement for Complex Reading Comprehension Tests under Generalizability Theory Models

G. Lee

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

**Guemin Lee**
**CTB/McGraw-Hill**

**Paper Presented at the Annual Meeting**
**of the American Educational Research Association**
**New Orleans, LA**
**April 24, 2000**

TM031278

BEST COPY AVAILABLE

**Abstract**

The purpose of this study was to investigate the relative appropriateness of several procedures for estimating reliability and standard errors of measurement of complex reading comprehension tests. Seven generalizability theory models were conceptualized by incorporating one or several factors of items, passages, themes, contents, and types of passages as sources of score variation. Results indicated that generalizability (reliability-like) coefficients for multivariate generalizability theory models incorporating "contents" and "types of passages" are close to coefficient alpha and, in contrast, incorporating "passages" and "themes" within univariate generalizability theory models produce non-negligible differences in reliability from coefficient alpha. This suggests that passages and themes be considered in evaluating the reliability of test scores for complex reading comprehension tests.

# Estimating Reliability and Standard Error of Measurement for Complex Reading Comprehension Tests Under Generalizability Theory Models

Previous studies have indicated that the reliability of test scores from reading comprehension tests (composed of passages and corresponding groups of items) is overestimated by conventional item-based reliability estimation methods (Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer & Thissen, 1996; Lee & Frisbie, 1999; Lee, 2000). Sireci, Thissen, and Wainer (1991) studied this topic using Bock's (1972) nominal model and concluded that the overestimation is due to "local dependence" among within-passage items. Lee and Frisbie (1999), using the person ($p$) by item ($i$) nested within passage ($h$) generalizability study design [ $p \times (i : h)$ ], provided reasons for the overestimation when coefficient alpha is used and contemplated the factors influencing the magnitude of the overestimation.

These studies have focused on only the dependence among items within passages. Other factors such as themes, contents, and types of passages were not considered. Little is known about how these variables affect estimates of reliability and standard error of measurement. This study had three primary objectives:

1. Estimate reliability and standard error of measurement for complex reading comprehension tests under various univariate and multivariate generalizability theory models.
2. Determine the magnitude of bias from using coefficient alpha in estimating reliability for test scores instead of using each of the generalizability theory approaches.
3. Investigate the influence of passage, contents, types of passages, and themes effects on the reliability of test scores from complex reading comprehension tests.

## Generalizability Theory Models

Seven generalizability theory models were conceptualized in this study. They considered factors such as items, passages, themes, contents, and/or types of passages as sources of score variation.

Univariate Generalizability Theory Model: $p \times i$

This design is the simplest one in that it identifies items as a unique source of error variation. Other sources of score variation such as passage, themes, contents, and types of passages are ignored in this design. The generalizability coefficient (reliability-like coefficient) of the pxI random effects decision study[1] design produces exactly the same value as coefficient alpha when the same measurement procedures are specified in a D-study as those used in the actual testing.

The univariate $p \times i$ generalizability study[2] design, persons ($p$) crossed with items ($i$), is appropriate for estimating variance components for this situation. The linear model for the response of a person to an item treats persons as objects of measurement and items as a random facet. The linear model can be represented as

$$X_{pi} = \mu \sim + \mu_p \sim + \mu_i \sim + \mu_{pi,e} \sim , \qquad (1)$$

where the terms of right-hand side are the grand mean, person effect, item effect, and person by item interaction effect confounded with unexplained sources of error, respectively.

Univariate Generalizability Theory Model: $p \times (i : h)$

It is well known that reading comprehension tests are composed of passages and corresponding groups of items. Several items are dependent upon some passages. The univariate $p \times (i : h)$ generalizability study design, persons ($p$) crossed with items ($i$) nested within passages ($h$), is appropriate for estimating variance components for this situation. The linear model for the response of a person to an item within a passage treats persons as objects of measurement and items and passages as random facets. This linear model can be represented as

$$X_{pih} = \mu \sim + \mu_p \sim + \mu_{i:h} \sim + \mu_h \sim + \mu_{ph} \sim + \mu_{pi:h,e} \sim . \qquad (2)$$

---

[1] Decision study (D-study) is a study conducted for the purpose of determining the most efficient measurement procedures for a given situation. It involves gathering data to inform a decision.

[2] Generalizability study (G-study) is done to determine how generalizable the scores can be for multiple situations. A G-study involves estimating variance components that might in turn be used in a D-study.

where the terms of right-hand side are the grand mean, person effect, item within passage effect, passage effect, person by passage interaction effect, and person by item within passage interaction effect confounded with unexplained sources of error, respectively.

Univariate Generalizability Theory Model: $p \times (i:h:t)$

In addition to items and passages, in some reading comprehension tests, "themes" may be introduced for grouping several passages and groups of items. For example, a reading comprehension test may be composed of two themes, "sports" and "machines", and four passages are related to the "sports" theme and five passages are connected to the "machines" theme. Consequently, the reading comprehension test is divided into two parts in this case, and several introductory statements can be given in front of each part for explaining the general idea about the theme.

The univariate $p \times (i:h:t)$ generalizability study design, persons ($p$) crossed with items ($i$) nested within passages ($h$) nested within themes ($t$), is appropriate for estimating variance components in this situation. The linear model for the response of a person to an item within a passage nested within a theme treats persons as objects of measurement and items, passages, and themes as random facets. The linear model can be represented as

$$X_{piht} = \mu \sim +\mu_p \sim +\mu_{i:h:t} \sim +\mu_{h:t} \sim +\mu_t \sim +\mu_{pt} \sim +\mu_{ph:t} \sim +\mu_{pi:h:t,e} \sim , \quad (3)$$

where the terms on the right-hand side are the grand mean, person effect, item within passage nested within theme effect, passage within theme effect, theme effect, person by theme interaction effect, person by passage within theme interaction effect, and person by item within passage nested within theme interaction effect confounded with unexplained sources of error, respectively.

Multivariate Generalizability Theory Model: $p \times i|C$

Usually, tests are constructed by following a table of specifications. In this case, items are written to sample each of several content strata, which are specified in the table of specifications. Stratified coefficient alpha was originally developed for this situation (Cronbach, Schönenmann, & McKie, 1965). The multivariate $p \times i|C$ generalizability study design, persons ($p$) crossed with items ($i$) for each content

BEST COPY AVAILABLE

stratum ($C$), is appropriate for estimating variance components in this situation. The linear model for the response of a person to an item for each content stratum treats persons as objects of measurement and items as a random facet. This linear model can be represented as

$$_{c}X_{pi} = {}_{c}\mu + {}_{c}\mu_{p} + {}_{c}\mu_{i} + {}_{c}\mu_{pi,e}, \tag{4}$$

for each content stratum. The terms on the right-hand side are the grand mean, person effect, item effect, and person by item interaction effect confounded with unexplained sources of error, respectively, for each content stratum.

Multivariate Generalizability Theory Model: $p \times i | M$

As Feldt and Brennan (1989) indicated, a reading comprehension test includes passages of several different types. There might be a poem, a short essay, an excerpt from a novel, some dialogue from a play, a newspaper article, and so on. It is reasonable to expect that parallel forms of a reading comprehension test include one or two passages from pre-specified types of passages. The multivariate $p \times i | M$ generalizability study design, persons ($p$) crossed with items ($i$) for each type of passage ($M$), is appropriate in this situation. The linear model is the same as Equation 4 except that the fixed facet is the types of passage ($M$) instead of the content strata ($C$).

Multivariate Generalizability Theory Model: $p \times (i:h) | C$

This design is different from the $p \times i | C$ design in that this design involves passages as well as items as random facets for each content stratum. That is, passages are assumed randomly sampled from a universe of passages, and items are assumed randomly sampled from that passage for each content stratum. The multivariate $p \times (i:h) | C$ generalizability study design, persons ($p$) crossed with items ($i$) nested within passages ($h$) for each content stratum ($C$), is appropriate for estimating variance components for this situation. The linear model for the response of a person to an item within a passage treats persons as objects of measurement and items and passages as random facets. This linear model can be represented as

$$_{c}X_{pih} = {}_{c}\mu + {}_{c}\mu_{p} + {}_{c}\mu_{i:h} + {}_{c}\mu_{h} + {}_{c}\mu_{ph} + {}_{c}\mu_{pi:h,e}, \tag{5}$$

for each content stratum. The terms on the right-hand side are the grand mean, person effect, item within passage effect, passage effect, person by passage interaction effect, and person by item within passage interaction effect confounded with unexplained sources of error, respectively, for each content stratum.

Multivariate Generalizability Theory Model: $p \times (i:h)|M$

This design is different from the $p \times i|M$ design in that this design assumes that passages as well as items are randomly sampled. In a generalizability framework, passages are assumed randomly sampled from a universe of passages within specified types of passages and items are assumed randomly sampled from that passage for each type of passage. The multivariate $p \times (i:h)|M$ generalizability study design, persons ($p$) crossed with items ($i$) nested within passages ($h$) for each type of passages ($M$), is appropriate in this situation. The linear model is the same as Equation 5 except that the fixed facet is the types of passage ($M$) instead of the content strata ($C$).

## Methods

Instruments

Several reading comprehension tests in achievement test batteries were used in the current study as an example of complex reading comprehension tests. Some items in the those reading comprehension tests focus on the central meaning of a passage rather than on surface details. Items cover various aspects of cognitive skills from initial understanding through development of interpretation and extension of concepts to other contexts. In addition to comprehension-type items, language usage questions are asked within the context of reading passages. The specific objectives and item allocation are presented in Table 1.

---

Insert Table 1 About Here

---

The majority of reading passages are taken from published work. Among the reading selections are excerpts from traditional and contemporary literature, informational selections from current publications, and real-life documents and graphics. Two test development experts classified these passages

into five categories – fiction, poetry, narrative article, document, and interview. Fiction referred to contemporary stories and traditional fables or myths, usually excerpted from published works. Poetry referred to short or long poems from published authors. Narrative articles were continuous prose based on facts, including biographies, autobiographies, magazine articles, and essays. Documents were reading materials presented in a graphic format such as maps, charts, tables, and forms used in school and work. Interviews referred to passages containing factual information gathered from talking to an individual, which were presented in a question-and-answer format.

Reading selections in reading comprehension tests used in this study are further characterized by the use of themes. Themes provide a framework supporting the assessment and connections that link the passages while permitting a range of styles, formats, and subjects for students to explore. That is, reading passages and corresponding question sets in the test are linked by broad themes designed to appeal to the age group being tested. Each theme is briefly described in an introduction that serves to elicit interest and orient students to the tasks ahead. Table 2 shows the themes, types of passages, and associated passages and items.

---

Insert Table 2 About Here

---

Data Sources

Data sets for the Reading Comprehension tests from students in grades 8 and 10 were used. The sample sizes were 2,114 for grade 8 and 1,351 for grade 10. The Reading Comprehension tests for both grades are composed of two or three parts related to the themes. There are 48 multiple choice items for both grades. The sample sizes and the general characteristics of each test are presented in Table 3.

---

Insert Table 3 About Here

---

Analyses

Generalizability analyses were conducted to estimate variance components. Because the number of items per passage usually varied, the conditions for a balanced design were not usually met in the

Reading Comprehension tests (Lee & Frisbie, 1999; Brennan, Jarjoura, & Deaton, 1980; Jarjoura & Brennan, 1981). Consequently, ANOVA-like procedures were used with urGENOVA (Brennan, 1999b) computer application program to estimate variance components for an unbalanced design. For the multivariate generalizability study treating either content strata or types of passages as a fixed facet, mGENOVA (Brennan, 1999a) application program was used for estimating variance components. Coefficient alphas and standard errors of measurement were computed to compare their values to the generalizability coefficient and standard error of measurement estimated from each generalizability theory model.

## Results and Discussion

Comparison of G-coefficients and SEMs

Table 4 provides generalizability coefficients (G-coefficients) and standard errors of measurement (SEMs) based on the several generalizability theory models. The pxI|M design produced the highest G-coefficients and the smallest SEMs in both Grades 8 and 10. However, the estimates of G-coefficient and SEM for the pxI|M design were similar to those from the pxI and pxI|C designs. In contrast, the px(I:H:T) design provided much lower G-coefficients and much larger SEMs for both grades, especially in the Grade 10, than did other designs.

---

Insert Table 4 About Here

---

Two points should be considered that help us understand general characteristics and tendencies in G-coefficients and SEMs. First, if more facets are incorporated within univariate generalizability frameworks, more error sources can be identified and, consequently, lower G-coefficients and larger SEMs can be expected (Lee & Frisbie, 1999). This argument can be supported by the results of the current study from comparison of G-coefficients and SEMs between the pxI and px(I:H) designs and comparison of those between the px(I:H) and px(I:H:T) designs.

Second, incorporating fixed facets within multivariate generalizability frameworks would produce higher G-coefficients and smaller SEMs. This argument can be confirmed by comparison between the pxI and pxI|C (or pxI|M) designs and comparison between the px(I:H) and px(I:H)|C (or px(I:H)|M) designs.

Based upon these two generalizations, it seems logical to expect some orders of G-coefficients (or reverse orders for SEMs) in terms of inequalities:

a. $px(I:H:T) < px(I:H) < pxI < pxI|C$ or $pxI|M$

b. $px(I:H:T) < px(I:H) < px(I:H)|C$ or $px(I:H)|M < pxI|C$ or $pxI|M$

The results from the current study support this kind of expectation.

Based only on the two considerations, it is difficult to anticipate inequality between the pxI and px(I:H)|C or between the pxI and px(I:H)|M designs. That is, in both px(I:H)|C and px(I:H)|M) designs, the passage facet was incorporated within an univariate framework and the content strata or types of passage facet was incorporated as a fixed facet within a multivariate generalizability framework. Thus, there should be compensation between random facets such as passages and fixed facets such as contents or types of passages. However, observed results of this study indicated that the passage effect was more influential on the size of G-coefficients than effects of content strata or types of passages.

Comparison with coefficient alpha

Coefficient alpha is a popular formula used to estimate reliability for a set of test scores. Coefficient alpha identifies items as a unique source of error. Consequently, it probably oversimplifies measurement procedures and leads to biased estimates for reliability and standard errors of measurement for the complex reading comprehension tests. Because coefficient alpha is widely used, it is meaningful to compare G-coefficients from various generalizability theory models with coefficient alpha. The differences between various G-coefficients and coefficient alpha are presented in Figure 1.

---

Insert Figure 1 About Here

---

Figure 1 shows that coefficient alpha was very similar to the G-coefficients from the pxI|C and pxI|M designs in both grades and also to the px(I:H)|M in grade 8. This implies that incorporating

"contents" or "types of passages" facet does not make any significant difference in reliability estimates. However, coefficient alpha was somewhat different from the G-coefficient for the px(I:H) design. That is, incorporating passage facet in addition to item facet made some non-negligible difference in reliability estimates. The difference between coefficient alpha and the G-coefficient was more evident when themes were considered as well as items and passages. In grade 10 reading comprehension test, the difference between the G-coefficient for the px(I:H:T) design and coefficient alpha was about -0.1. This difference seems big enough from a practical standpoint to suggest that "passages" and "themes" be considered when one is evaluating the reliability of a set of test scores for complex reading comprehension tests.

Passage Effects

The differences of G-coefficients between the pxI and px(I:H) designs were 0.022 for grade 8 and 0.037 for grade 10. The results are consistent with Lee and Frisbie (1999) even though the magnitudes of differences between the pxI and px(I:H) designs are somewhat different. They reported a little bigger difference for grade 8 (0.040 difference) and similar difference for grade 11 (0.034 difference). As Lee and Frisbie (1999) indicated, the person by passage interaction variance component in a D-study, $\hat{\sigma}^2(pH)$, contributes to the universe score variance, analogous to true score variance, in the pxI design, but it contributes to the error score variance in the px(I:H) design. Consequently, the G-coefficient from the pxI design is greater than that from the px(I:H) design. The reliability estimation methods ignoring passage facet lead to positively biased estimates for reliability for test scores involving passages. Thus, the difference of G-coefficients between the pxI and px(I:H) designs would be related to the magnitude of $\hat{\sigma}^2(pH)$, the variance component estimate for the person by passage interaction effect. The variance component estimates and G-coefficient differences between the pxI and px(I:H) designs are presented in Table 5.

---

Insert Table 5 About Here

---

Content Strata and Types of Passage Effects

Content strata and types of passages are treated as fixed factors in the current study. Whether a factor is random or fixed in a particular situation would depend on the sampling plan used to form the test (Lee, Dunbar, & Frisbie, 1999). In this case, the content strata (or types of passages) were not sampled from a universe of content strata (or a universe of types of passages). Because the contents (or types of passages) are replicated from form to form, this factor should be treated as fixed. In order to incorporate content strata or types of passages as a fixed facet, the multivariate generalizability frameworks were administered (Brennan, 1992, 1999a).

The differences of G-coefficients between the pxI and pxI|C designs were 0.000 for grade 8 and 0.001 for grade 10 and the differences between the pxI and pxI|M designs were 0.002 for grade 8 and 0.003 for grade 10. These differences seem too small to be considered meaningful for G-coefficients from a practical standpoint. These negligible differences can be explained by the substantial covariation among contents or among types of passages. For example, if each content stratum (or each type of passages) has perfect relations with other content strata (or other types of passages), it is unnecessary to differentiate distinct contents (or types of passages). In this special case, the pxI and pxI|C (or pxI and pxI|M) designs will provide the same G-coefficients and SEMs under an assumption of non-random errors for the estimates. To check this argument, observed correlations and disattenuated correlations among contents and among types of passages are computed and presented in Tables 6 and 7, respectively.

---

Insert Table 6 About Here

---

---

Insert Table 7 About Here

---

The disattenated correlations can be understood as correlations between the universe scores, analogous to true scores, for two contents or for two types of passages. High diattenuated correlations were found. Thus, it is logical to anticipate high level of agreement in G-coefficients between the pxI and pxI|C (or pxI and pxI|M) designs. The disattenuated correlations among contents were higher than those among types of passages. This might be used as one piece of evidence to explain slightly larger difference of G-

coefficients between the pxI and pxI|M designs than that between the pxI and pxI|C designs. For the grade 8 case, the disattenuated correlations among contents are almost 1 and both the pxI and pxI|C designs provided the same G-coefficients and SEMs.

Theme Effects

The differences in G-coefficients between the pxI and px(I:H:T) designs were 0.042 for grade 8 and 0.096 for grade 10. The differences of G-coefficients between the px(I:H) and px(I:H:T) designs were 0.020 for grade 8 and 0.059 for grade 10. These differences seem big enough that the theme facet should be considered in assessing the reliability of test scores for complex reading comprehension tests.

To examine the influence of themes on the reliability estimates of the px(I:H:T) random effects design, several D-studies were completed. In conducting several D-studies, the total number of items and the total number of passages were set to 48 and 9, respectively. In both grades, these numbers were the same as those used in the actual tests and the number of themes was varied from 1 to 9. The G-coefficients of the px(I:H:T) random effects D-study designs with varying number of themes are presented in Figure 2.

---

Insert Figure 2 About Here

---

Because the total number of items and total number of passage were fixed, varying the number of themes does not greatly impact testing time and any of the testing conditions. For example, if two themes were used, the first four passages might be related to the first theme and the following five passages might be related to the second theme. Assuming the use of one more theme, for a total of three themes, the first three passages might be related to the first theme, the next three passages to the second theme, and the last three passages might be related to the third theme. In both cases, because the total number of passages and items are the same, there is no need to change testing time.

A non-negligible increment of G-coefficients was found as the number of themes increased. Based upon the results, at lease three or four themes would be recommend to be used in a test for getting more accurate inference about students' ability scores. In a practical test construction situation, a graph like Figure 2 can be used to determine efficient measurement procedures. For example, in the grade 10 reading

comprehension test, when 0.86 is the desired level of reliability, about three themes are needed given the presence of 9 passages and 48 items.

## Conclusions

Three main generalizations follow from the findings of this study.

First, generalizability theory models incorporating more random facets within univariate generalizability frameworks produce lower generalizability coefficients and larger standard errors of measurement because they identify more sources of error. In contrast, generalizability theory models incorporating fixed facets within multivariate generalizability frameworks produce higher generalizability coefficients and smaller standard errors of measurement.

Second, generalizability coefficients that incorporate "contents" or "types of passages" within multivariate generalizability theory models produce values close to coefficient alpha. However, the use of generalizability theory models incorporating "passages" and "themes" within univariate generalizability frameworks results in some non-negligible differences in reliability estimates relative to coefficient alpha.

Third, the results of the current study suggest that the passages and themes facets be considered in evaluating the reliability of test scores for complex reading comprehension tests. Thus, the px(I:H:T), person crossed with items within passages nested within themes, design appears to be the most appropriate model among seven models conceptualized in this study

**References**

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Brennan, R.L. (1992). Elements of generalizability theory. Iowa City, IA: Am,erican College Testing.

Brennan, R.L. (1999a). Manual for mGENOVA (version 2.0). (Iowa Testing Programs Occasional Paper No. 47). Iowa City, IA: University of Iowa.

Brennan, R.L. (1999b). Manual for urGENOVA (version 1.4). (Iowa Testing Programs Occasional Paper No. 46). Iowa City, IA: University of Iowa.

Brennan, R.L., Jarjoura, D., & Deaton, E.L. (1980). Some issues concerning the estimation and interpretation of variance components in generalizability theory (ACT Tech. Bulletin No. 36). Iowa City, IA: American College Testing.

Cronbach, L.J., Schonenmann, P., & Mckie, D. (1965). Alpha coefficients for stratified-parallel tests. Educational and Psychological Measurement, 25, 291-312.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 105-146). Washington, DC: American Council on Education.

Jarjoura, D., & Brennan, R.L. (1981). Three variance components models for some measurement procedures in which unequal numbers of items fall into discrete categories (ACT Tech. Bulletin No. 37). Iowa City, IA: American College Testing.

Lee, G. (2000). Estimating conditional standard errors of measurement for tests composed of testlets. Applied Measurement in Education, 13, 161-180.

Lee, G., Dunbar, S.B., & Frisbie, D.A. (1999, April). Measurement models for a testlet-based test. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Lee, G., & Frisbie, D.A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. Applied Measurement in Education, 12, 237-255.

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28, 237-247.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. Applied Measurement in Education, 8, 157-186.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? Educational Measurement : Issues and Practice, 15, 22-29.

## Note

The author appreciates the assistance of Cara Davis-Jacobson and Melinda Manlin in classifying types of testlets for complex reading comprehension tests. The comments of Anne Fizpatrick and Robert Sykes are gratefully acknowledged.

**TABLE 1**

Objectives/Skills and Item Allocation for Reading Comprehension Tests

| Objective/Skill | Grade 8 | Grade 10 |
|---|---|---|
| Basic understanding | 13 (27.1) | 11 (22.9) |
| - vocabulary | | |
| - stated information | | |
| - stated information graphics | | |
| Analyze text | 16 (33.3) | 14 (29.2) |
| - main idea/theme | | |
| - supporting evidence | | |
| - conclusions | | |
| - cause/effect | | |
| - story elements/plot | | |
| - story element/character | | |
| - literary techniques | | |
| - nonfiction elements | | |
| Evaluate and extend meaning | 8 (16.7) | 10 (20.8) |
| - author/purpose | | |
| - author/point of view | | |
| - author/tone | | |
| - predict/hypothesize | | |
| - extend/apply meaning | | |
| - critical assessment | | |
| Identify reading strategies | 11 (22.9) | 13 (27.1) |
| - make connections | | |
| - apply genre criteria | | |
| - utilize structure | | |
| - vocabulary strategies | | |
| - self-monitor | | |
| - graphic strategies | | |

Note. The number in the parenthesis represents the percentage of items in a test.

**TABLE 2**

Themes, Types of Passages, and Item Allocation in Reading Comprehension Tests

| Theme | Passage Number | Type of Passage | No. of Items per Passage |
|---|---|---|---|
| | | Grade 8 | |
| 1. Challenges | 1 | Fiction | 7 |
| | 2 | Fiction | 5 |
| | 3 | Document | 2 |
| | 4 | Narrative Article | 4 |
| 2. Universe | 5 | Poetry | 5 |
| | 6 | Interview | 8 |
| | 7 | Document | 4 |
| 3. World of Work | 8 | Narrative Article | 8 |
| | 9 | Narrative Article | 5 |
| Total | 9 Passages | | 48 Itesms |
| | | Grade 10 | |
| 1. Flight | 1 | Fiction | 10 |
| | 2 | Narrative Article | 3 |
| | 3 | Document | 3 |
| | 4 | Narrative Article | 7 |
| 2. Bones | 5 | Interview | 9 |
| | 6 | Interview | 5 |
| | 7 | Document | 2 |
| | 8 | Interview | 6 |
| | 9 | Document | 3 |
| Total | 9 Passages | | 48 Itesms |

**TABLE 3**
Descriptive Statistics for Data Sources Used in This Study

| | Grade 8 Reading Comprehension | Grade 10 Reading Comprehension |
|---|---|---|
| Sample size | 2,114 | 1,351 |
| Raw Score Mean | 33.6 | 31.7 |
| Raw Score Standard Deviation | 10.32 | 10.94 |
| Raw Score Skewness | -0.562 | -0.428 |
| Raw Score Kurtosis | 2.156 | 2.053 |

**TABLE 4**

Generalizability Coefficients and Standard Errors of Measurement Based on the Several Generalizability Theory Models for Reading Comprehension Tests

| Model | No. of Random | No. of Fixed | Grade 8 G-Coefficient | Grade 8 SEM | Grade 10 G-Coefficient | Grade 10 SEM |
|---|---|---|---|---|---|---|
| pxI | 1 | 0 | 0.932 | 2.694 | 0.934 | 2.803 |
| px(I:H) | 2 | 0 | 0.910 | 3.092 | 0.897 | 3.508 |
| pxI \| C | 1 | 1 | 0.932 | 2.694 | 0.935 | 2.791 |
| pxI \| M | 1 | 1 | 0.934 | 2.647 | 0.937 | 2.741 |
| px(I:H:T) | 3 | 0 | 0.890 | 3.408 | 0.838 | 4.391 |
| px(I:H) \| C | 2 | 1 | 0.921 | 2.955 | 0.904 | 3.553 |
| px(I:H) \| M | 2 | 1 | 0.929 | 2.745 | 0.920 | 3.088 |

Note. No. of Random = number of random facets; No. of Fixed = number of fixed facets; G-Coefficient = generalizability coefficient; SEM = standard error of measurement.

**TABLE 5**

Variance Component Estimates for the Random Effects px(i:h) Generalizability Theory Model for Reading Comprehension Tests

| Variance Component/ G-Coeff. | Lee & Frisbie (1999) | | Current Study | |
|---|---|---|---|---|
| | Grade 8 | Grade 11 | Grade 8 | Grade 10 |
| $\hat{\sigma}^2(p)$ | 4.1 | 4.8 | 4.2 | 4.7 |
| $\hat{\sigma}^2(h)$ | 0.6 | 0.1 | 0.0 | 0.5 |
| $\hat{\sigma}^2(i:h)$ | 1.1 | 1.0 | 1.6 | 0.8 |
| $\hat{\sigma}^2(ph)$ | 1.6 | 1.0 | 0.9 | 1.6 |
| $\hat{\sigma}^2(pi:h)$ | 17.6 | 16.7 | 14.3 | 15.0 |
| pxI G-Coeff. (a) | 0.928 | 0.926 | 0.932 | 0.934 |
| px(I:H) G-Coeff. (b) | 0.888 | 0.892 | 0.910 | 0.897 |
| Difference (a-b) | 0.040 | 0.034 | 0.022 | 0.037 |

Notes. The scale of the variance component estimates was changed by multiplying all entries by 100 and then rounding to one decimal place. G-Coeff. = generalizability coefficient.

**TABLE 6**

Observed and Disattenuated Correlations among Contents in Reading Comprehension Tests for Grades 8 and 10

| | Basic Understanding (BU) | Analyze Text (AT) | Evaluate/Extend Meaning (EM) | Identify Reading Strategies (IS) |
|---|---|---|---|---|
| | | Grade 8 | | |
| BU | | 1.010 | 1.009 | 1.002 |
| AT | 0.811 | | 1.019 | 0.997 |
| EM | 0.736 | 0.741 | | 1.006 |
| IS | 0.796 | 0.790 | 0.724 | |
| | | Grade 10 | | |
| BU | | 0.980 | 0.992 | 0.956 |
| AT | 0.759 | | 1.008 | 0.960 |
| EM | 0.756 | 0.777 | | 0.932 |
| IS | 0.761 | 0.772 | 0.740 | |

Note. Lower-diagonal elements are observed correlations and upper-diagonal elements are disattenuated correlations.

**TABLE 7**
Observed and Disattenuated Correlations among Types of Passages in Reading Comprehension Tests for Grades 8 and 10

| | Fiction | Poetry | Narr. Article | Document | Interview |
|---|---|---|---|---|---|
| | | | Grade 8 | | |
| Fiction | | 0.899 | 0.890 | 0.907 | 0.837 |
| Poetry | 0.600 | | 0.879 | 0.870 | 0.924 |
| Narr. Article | 0.711 | 0.634 | | 0.905 | 0.874 |
| Document | 0.605 | 0.524 | 0.652 | | 0.843 |
| Interview | 0.636 | 0.634 | 0.717 | 0.578 | |
| | | | Grade 10 | | |
| Fiction | | N/A | 0.844 | 0.775 | 0.760 |
| Poetry | N/A | | N/A | N/A | N/A |
| Narr. Article | 0.652 | N/A | | 0.819 | 0.812 |
| Document | 0.577 | N/A | 0.636 | | 0.922 |
| Interview | 0.613 | N/A | 0.683 | 0.748 | |

Note. Lower-diagonal elements are observed correlations and upper-diagonal elements are disattenuated correlations

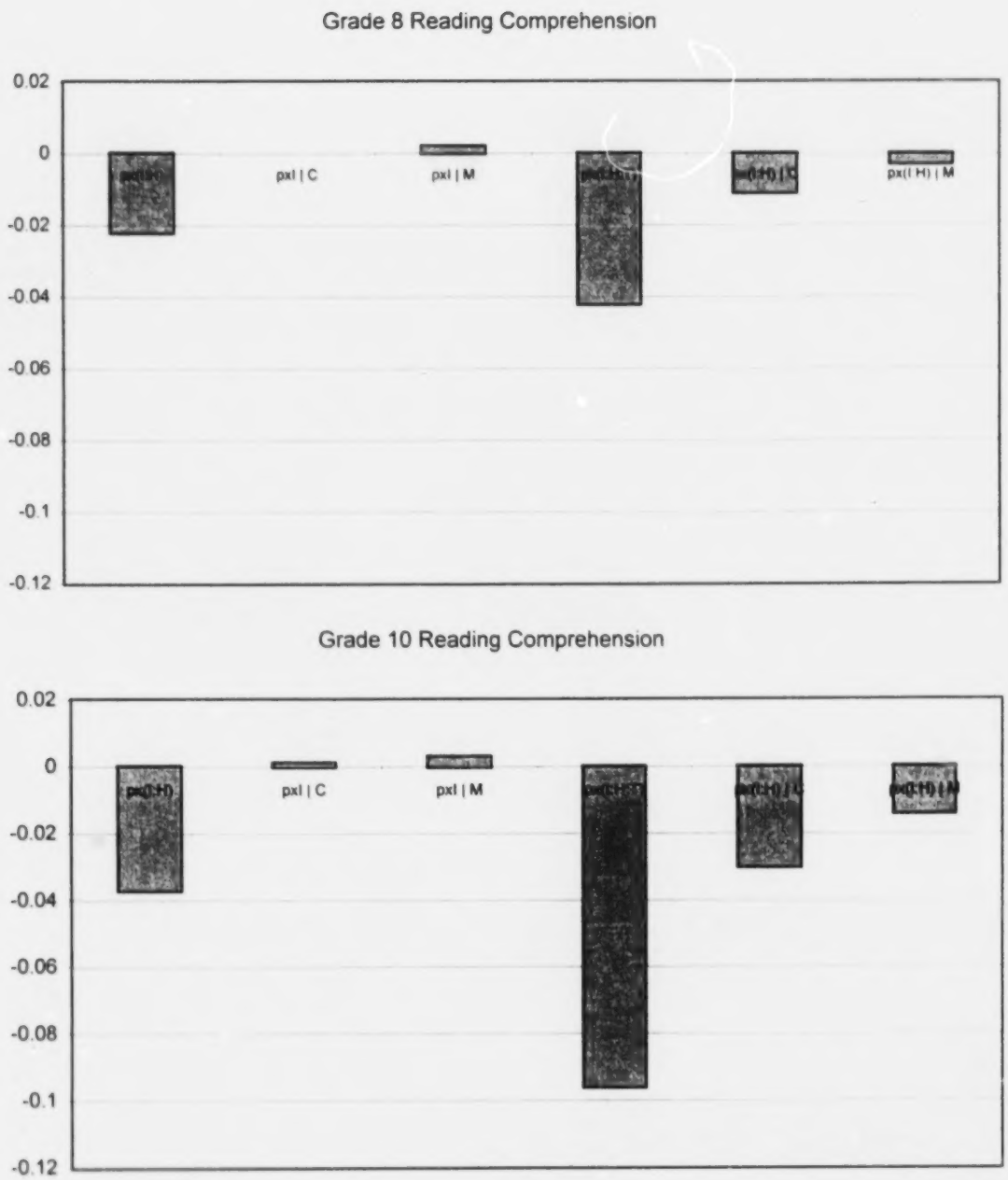


Figure 1. Difference between generalizability coefficients and coefficient alpha using coefficient alpha as a baseline



BEST COPY AVAILABLE

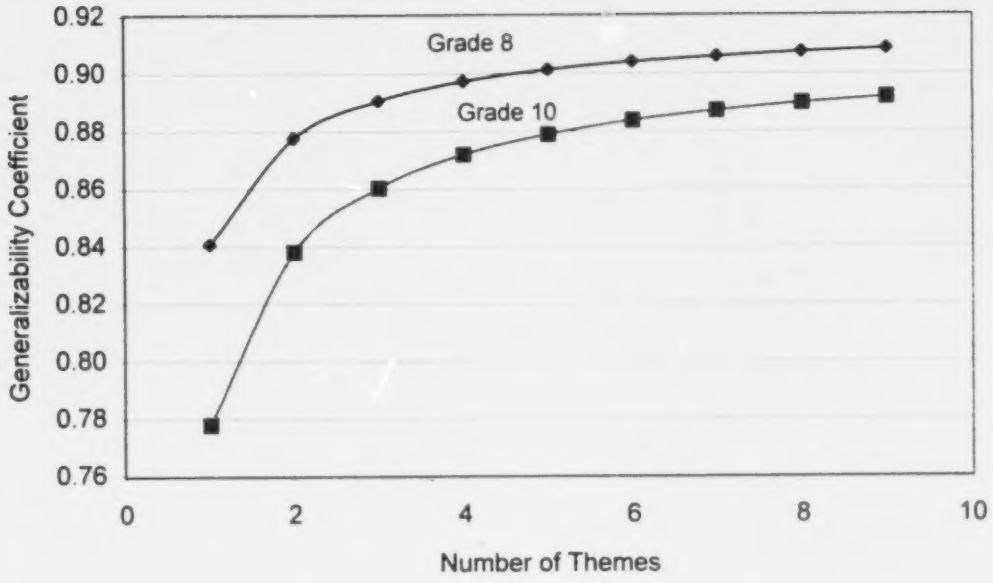


Figure 2. The theme effects on generalizability coefficients for given test length.